The VIMSS Computational Microbiology Core (http://escalante.lbl.gov)

Eric Alm[1], Katherine Huang[1], Morgan Price[1], Richard Koche[1,2], Yue Wang[1,2], Inna Dubchak[1,3], Adam Arkin[1,2]

1. Lawrence Berkeley National Lab, Berkeley, CA 2. University of California, Berkeley, Bioengineering department, Berkeley,CA 3. Joint Genome Institue, Walnut Creek, CA

Virtual Institute of Microbial Stress and Survival

Introduction

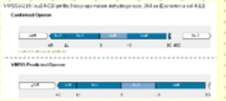
The primary roles of the VIMSS Computational Microbiology Core are to curate, analyze, and ultimately build models of data generated by the Applied Environmental Microbiology and Functional Genomics Core groups. The near-term focus of the computational group has been to build the scientific and technical infrastructure to carry out these roles. Central to each of these goals has been the development of a comprehensive relational database (VIMSSDB) that integrates genomic data and analyses together with data obtained from experiment.

VIMSS Comparative Genomics Database

VIMSSDB. At present, well over 100 microbial genomes have been sequenced, and hundreds more are currently in the pipeline. Despite this fact, tools to explore this wealth of information have focused on individual genome sequences. The VIMSS Comparative Genomics database and web-based tools are designed to facilitate cross-species comparison, as well as to integrate experimental data sets with genome-scale functional annotations such as operon and regulon predictions, metabolic maps, and gene annotations according to the Gene Ontology. Over 130 complete genome sequences are represented in the VIMSS Comparative Genomics Database, which is implemented as a MySQL relational database, a Perl library for accessing the database, and a user-friendly website designed for laboratory biologists.

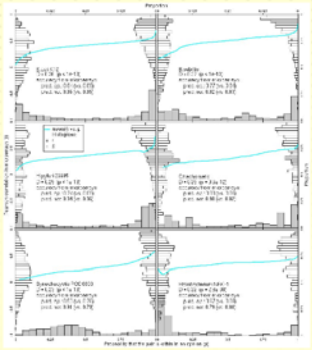
Predicting Operons in All Prokaryotes

Operon Prediction. Operons are the fundamental unit of transcriptional regulation in prokaryotes, yet little is known about operon structure outside a few model organisms. To identify operons in other prokaryotes, we combine predictions inferred from conservation of gene order across 129 prokayotic genomes with functional prediction based on sequence homology and use these to infer a genome-specific model of the intergenic distances between adjacent pairs genes on the same operon and pairs that span a transcriptional boundary. We combine these comparative genomics scores with our distance-based score to make predictions for 129 genomes. To validate these predictions, we compare against microarray data for six diverse prokaryotes: *Escherichia coli, Bacillus subtilis, Helicobacter pylori, Synechocystis sp. PCC6803, Chlamydia trachomatis, and the archeon Halobacterium sp. NRC-1.* We conclude that our genome-specific distances models are accurate and validate differences from the *E. coli* model using microarray data for *Helicobacter pylori* and *Halobacterium.* Further, we find that contrary to earlier reports, *H. pylori* has many operons, and that *Synechocystis* has a significant number of operons despite its unusual intergenic distances. Finally, we observe that genomes with the majority of their genes on the leading strand of replication have an even higher proportion of multigene transcripts on the leading strand, leading to an estimate of the total number of operons in these genomes.



VIMSS Operon Browser (Left). The VIMSS protein pages allows users to view known operons as well as predictions that contain the query gene. Operon predictions are available for 129 complete genome sequences.

Validating Predictions in Six Species Using Microarrays (Below). We use the similarity of microarray expression profiles to test our operon predictions for adjacent pairs of genes on the same strand of DNA. The cyan line shows the smoothed average of expression correlation (r, y-axis) for gene pairs with a given probability of being in the same operon (p, x-axis). The histograms on the left axes represent the distributions of correlations in expression profiles for gene pairs predicted to be different operons (p < 0.5). On the right axes are histograms of correlations in expression between genes predicted to be in the same operon (p > 0.5). The bottom axis shows the distribution of p-values obtained for predictions in each genome. Also shown are accuracy estimates based on the microarrays, as well as accuracy predicted by the model (in parentheses).

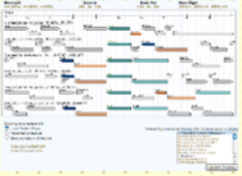


VIMSS Web Tools

VIMSS Comparative Genomics Web Tools. The VIMSS Genome Browser allows users to align any number of genomes and Identifies predicted orthology relationships between genes. From the browser, users can save genes of interest for use in the VIMSS Bioinformatics Workbench, explore individual genes in depth by clicking to the Protein Pages for information about functional annotations, sequence domains, BLAST alignments, predicted operon structure and functionally related genes inferred from a combination of comparative genomics methods and microarray experiments. The VertiGO comparative gene ontology browser allows users to simultaneously view the genetic complement of any number of genomes according to the Gene Ontology hierarchy. A metabolism browser based on the KEGG metabolic maps allows browsing either the set of enzymes predicted to be present in a single genome, or a comparison highlighting the metabolic differences between two genomes. The VIMSS Bioinformatics Workbench allows users to create and save lists of genes of interest, and use these lists to investigate phylogenetic relationships by making multiple sequence alignments and phylogenetic trees.

VIMSS Protein Pages and Genome Browser

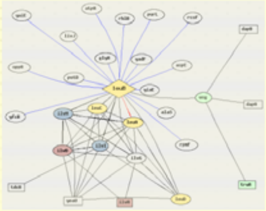
Gene Info Page (Above). The Gene Info page is the starting point for information about a gene of interest. The navigation bar at top allows users to quickly access operon & regulon predictions for that gene, domain alignments, homologs, and nt or aa sequence data. From this page users can also add their own annotation to a gene, save the gene for future study, or jump to the Genome Browser (shown below).

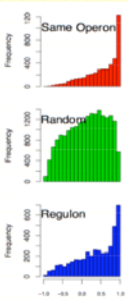


Comparative Genome Browser (Above). The Genome Browser displays orthologous regions from any number of species simultaneously. Predicted orthology relationships among genes are color coded on the display. From the browser, genes can be saved for later use in the Bioinformatics Workbench, or can be investigated in further depth by clicking links to the protein pages for each gene.

Predicting Regulons

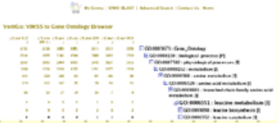
Comparative Genomics. Conserved gene order across distantly related taxa may indicate operon structure. Moreover, when that operon structure is disrupted in one species, the resulting transcriptional units may share similar regulatatory logic. A test of this hypothesis using gene expression microarray data is shown in the figure below (right). Below left is a screenshot of the VIMSS regulon browser, which allows users to browse the neghborhood of genes predicted to be coregulated (based on conserved gene order in distant genomes, black lines), or observed to be coregulated in microarray experiments (blue lines; red lines indicate connections both predicted and observed).





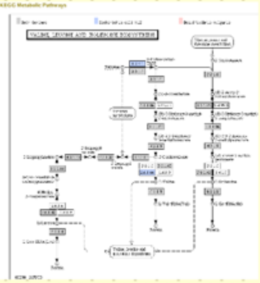
Microarrays Show Disrupted Operons Share Similar Regulation (Right). Pairs of *E. coli* genes from three categories were selected: (i) genes predicted to occur on the same transcript ("Operon" pairs), (ii) randomly selected pairs of genes ("Random" pairs), and (iii) genes that tend to occur in close proximity on the chromosome (and presumably on the same operon) in several distantly related taxa, but do not occur in the same operon in *E. coli* ("Regulon" pairs). Shown are the distributions of Pearson correlation coefficients for gene pairs in each of the categories. Predicted "Regulon" pairs are significantly and only slightly less correlated than predicted operon pairs.

Comparing Gene Content in Several Species

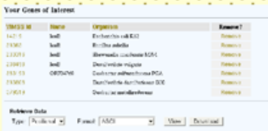


Comparative GO Browser (Left). The GO hierarchy is a convenient way to map functions onto genes using a controlled vocabulary of terms. The VIMSS GO browser allows users to browse through the GO hierarchy comparing the gene content in terms of GO annotations for any number of target genomes. Additional links take users to the protein pages for more detail on selected genes.

Comparative Metabolic Maps (Right). KEGG metabolic maps are available at the VIMSS site to either browse the metabolic capability of a single organism, or to highlight differences between twos organisms. Shown at right is a comparison of enzymes used by *E. coli* and *D. vulgaris* in the Val/Leu/Ile biosynthesis pathway.



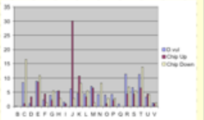
VIMSS Bioinformatics Workbench



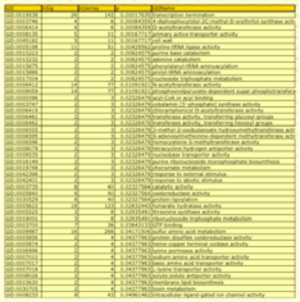
VIMSS Bioinformatics Workbench. Users of the comparative genomics web tools can select genes of interest from a number of locations on the VIMSS site. These genes are saved in the user's "Shopping Cart", which is shown above. These gene lists can be saved permanently, and used for further analysis within the VIMSS site. Shown below, genes selected from above are used to make a multiple sequence alignment and phylogenetic tree for various homologs of the *E. coli* leuB gene.

Functional Genomics Data Analysis

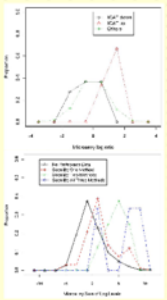
Data Analysis. The Computational Core group is responsible for statistical analysis of the large quantities of data being generated by the Functional Genomics Core group, and for integration of that data with genomic annotations and functional predictions.



Summary of O_2 Stress in *D. vulgaris* Using COG Functional Classes (Above). The distribution of COG functional categories for *D.* vulgaris genes predicted to be differentially expressed in response to O_2 stress is shown above. Red bars indicate up-regulated genes, yellow bars represent down-regulated genes, and blue bars represent total genes with that functional category assignment. COG assignments were made automatically using RPS-BLAST against the CDD database. Categories: B - Chromatin, C - Energy production, D - Cell cycle, E - AA metabolism, F - Nucleotide metabolism, G - Carbohydrate metabolism, H - Coenzyme metabolism, I - Lipid metabolism, J - Translation, K - Transcription, L - Replication, M - Cell wall, N - Motility, O - Posttranslational modifications, P - Ion transport/metabolism, Q - Secondary metabolites, R - General, S - Unknown, T - Signal transduction, U - Intracellular trafficking, V - Defense mechanisms



Using the GO Hierarchy to Interpret Gene Expression Experiments (Above). The Gene Ontology (GO) Consortium has developed a controlled vocabulary for gene functional annotation (*Ashburner M et al. Genome Res. 2001 Aug; 11(8): 1425-33*) that provides a quick and efficient way to identify clusters of coexpressed genes. It is especially useful for clustering gene expression data with a limited amount of independent experimental conditions. We performed the Fisher exact test on the GO categories to identify those terms with over-reperesented number of differentially expressed genes. Shown above are test results for gene expression studies in the anaerobe *D. vulgaris* after 2 hr. exposure to O_2. Shown for each high-scoring GO category are the GO term, the number of significant changers annotated with that term, the total number of genes matching the GO term with valid microarray spots, and a short description of each category. GO terms were assigned automatically based on InterPRO domain assignments.



Comparison of Proteomics and Transcriptomics. (Left). We compared microarray and proteomics data on *D. vulgaris* cells under similar conditions (O_2 stress) collected by the VIMSS Functional Genomics Core group. The top panel (left) compares the distribution of gene expression changes for three categories of genes: those predicted by ICAT measurements to be up-regulted, those predicted to be down-regulated, and those seen in the ICAT experiment, but not predicted with confidence to be differentially regulated. As expected, mRNA levels (taken after 2hrs of O_2), generally agree with proteomics results (taken after 5 hrs). The bottom panel (left) illustrates a potential limitation of proteomics methods. Data from three different mass spectrometry techniques are compared to total mRNA levels (estimated using genomic DNA as control). Genes identified by the proteomics techniques tend to be limited to the more highly expressed genes in the cell. Genes identified by two or more techniques show an even stronger trend toward highly expressed genes.